Audio Style Transfer to Improve Accent Parity in Automated Speech Recognition

Nicholas Chan
Faculty of Applied Science & Engineering
University of Toronto
Toronto, Canada
yticholas.chan@mail.utoronto.ca

Benjamin Mah
Faculty of Applied Science & Engineering
University of Toronto
Toronto, Canada
benjamin.mah@mail.utoronto.ca

William Yao
Faculty of Arts & Science
University of Toronto
Toronto, Canada
will.yao@mail.utoronto.ca

GitHub Repository: https://github.com/benjaminmah/accent-style-transfer

I. Introduction & Problem Statement

Automatic speech recognition (ASR) systems serve to transcribe human speech into text and have enabled the human voice to become a widespread means of interfacing with computers. However, the literature has shown that state-of-the-art (SOTA) English ASR models generally suffer from higher error rates when transcribing the speech of people with non-standard accents and vernaculars due to their lack of representation in most training datasets [1]. For instance, the Word Error Rates of commercial ASR systems on a corpus of African-American Vernacular English have previously been observed to be nearly 85 percent higher than those of the same systems on a corpus of Standard American English [2]. Our objective, therefore, is to increase parity in the transcription accuracy of ASR models between different varieties of the English language.

This project is of particular social impact due to the pressing importance of inclusive ASR systems. Speech recognition is currently used in many applications like voice assistants, closed captioning, and accessible data entry for which it is crucial to promote equitable use. Moreover, as ASRs are applied to increasingly impactful domains like courtroom and medical transcriptions, biases in these systems can create new avenues of technological disenfranchisement and discrimination against smaller language communities [3]. Although these issues are classically resolved by training ASR on a more diverse and representative training set, this solution is not trivial. The task of collecting large amounts of speech data can often be prohibitively expensive and time-consuming, especially considering the great geographical diversity with which different accents are dispersed worldwide.

In this paper, we propose a neural style transfer (NST) model that transforms audio inputs containing non-standard English accents into a standardized style more reminiscent of the accents on which SOTA ASRs are typically trained and perform better. The NST is achieved by training a convolutional neural network to learn features associated with the content and style of a speech sample input separately. We define content to be the words spoken in the sample, and style to be the pronunciation characteristics of those words specific to the speaker. By learning the style of a reference speaker whose accent is better recognized by existing ASR systems, and transferring this style to a target speech sample spoken with a non-standard accent, our model outputs an augmented form of the target sample that is more recognizable to the ASR system.

To evaluate our method, we pass the augmented target sample to a pre-trained ASR system and measure the system's accuracy on the sample. This is compared against passing the original target sample directly to the system as a baseline. Although our model does not demonstrate improvement in most of our empirical tests, there are promising hints that NST can be a viable method for this task given more computational resources and model adjustments. We hope that an extension of our approach will provide a generalizable and computationally inexpensive method of standardizing accents and improving the accessibility of SOTA ASR models for accented speakers.

II. Literature Review and Related Work

In "A Neural Algorithm of Artistic Style" by Gatys et al, the concept of separating and recombining content and style with deep neural networks is explored [4]. This paper proposes the use of convolutional neural networks to extract both content and style from an image; the content is stored in feature maps, where each feature map can be used to reconstruct the original image. For representing the image style, a separate feature space is built on top of the CNN to find the correlation between the feature maps from each layer. To demonstrate the separation of content and style, the paper combined the content from one image, with a style obtained from another image. Fig. 1. showcases the results of applying Vincent van Gogh's style to a regular image of houses.

The proposed model architecture includes 16 convolutional and 5 pooling layers. In addition, two loss functions (content loss and style loss) are used to train the model. The content loss is defined as the square-error loss between two feature representations, whereas the style loss is computed with the mean-squared distance between the feature correlations of the original and new image. When generating a new image, gradient descent can be applied to a white noise image, with

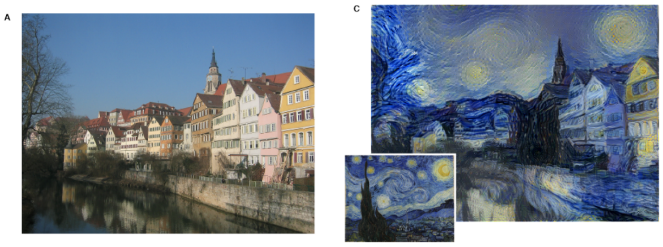

Fig. 1. Applying Vincent van Gogh's style to an image through style transfer. Adapted from [4]

the goal of minimizing the two losses. While this paper explores style transfer with two images, the same concept can be implemented for audio recordings such as voices.

Prior work by GitHub user mazzzystar explored voice style transfer with convolutional neural networks [5], with the aim of applying another person's voice to an existing voice recording for purposes such as musical remixes. This is accomplished by extracting the content from an original audio and the style from a voice recording by the person of interest. First, both recordings are converted to a spectrogram. Both images are then inputted to a CNN, which extracts the content and style from the respective spectrograms. Finally, a new spectrogram is constructed from a random noise image through gradient descent, which minimizes the content loss and style loss. The output spectrogram can then be converted back to an audio recording. This model does not require a large dataset, in addition to having a low training time. Our work primarily derives from the architecture presented in mazzzystar's work with changes made to orient the model towards accent transferring.

III. Methodology

A. Data Collection

For this study, we use the Speech Accent Archive prepared by Steven H. Weinberger of George Mason University. This dataset contains audio recordings from 2,140 individuals reciting the same English passage. Collectively, the speakers originate from 177 countries and have 214 native languages; as a result, a majority of the individuals speak English as a second language with a particular non-standard accent. Each individual recorded exactly one speech sample, so the dataset contains exactly 2,140 audio recordings. The text recited in each recording is intended to capture a wide variety of English phonemes and can be found below.

"Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station."

All speech samples in the Speech Accent Archive are stored in the .mp3 file format. Each file name is enumerated based on the primary language of the speaker. For instance, the file "english494.mp3" contains the sample recorded by the 494th individual in the dataset who speaks English as a first language.

B. Data Processing

In the data processing stage of our project pipeline, the primary objective is to convert raw audio from the native .mp3 format into a form suitable for input into the NST model. Using the Librosa library, the input audio file is decoded to extract its digital signal. This initial step is crucial for capturing the temporal dynamics and intrinsic properties of the audio content. We then use a combination of the Librosa and NoiseReduce libraries to resample the signal to 22050 Hz, peak normalize the signal, and reduce background noise, respectively. This data-cleaning step ensures that all speech samples passed to the model as input have consistent amplitude levels and relatively little random noise that may affect its performance. Following this, the decoded audio signals are transformed into spectrograms via the Short-Time Fourier Transform (STFT), a process that converts the time-domain signals into a frequency-domain representation. This provides both magnitude and phase information for each frequency component. The magnitude component of the representation is scaled logarithmically to improve the dynamic range and relevance of the spectrograms, ensuring subtle nuances in the audio are retained. The spectrograms are finally converted into tensors, making them suitable as inputs to the NST model. This preprocessing is necessary for the effective separation of the content and style features of the inputs, ensuring that the content's linguistic structure is maintained while the stylistic nuances are adopted from the target audio.

C. Model Architecture

The base of our proposed NST model is a convolutional neural network (CNN) for the purposes of feature extraction. The architecture processes input spectrograms with spatial dimensions $N \times M$, which represent the frequency bins by time frames, through two convolutional layers. These dimensions are preserved across layers to ensure a consistent feature extraction scale. The configuration, along with the specifics of kernel size, stride, and padding is represented in Table I.

TABLE I
Model Layer Specifications

Layer	Input Size	Output Size	Kernel Size	Stride	Padding
Input	[1, N, M]	-	-	-	-
Convolutional 1	[1, N, M]	[32, N, M]	(3, 5)	1	(1, 2)
Convolutional 2	[32, N, M]	[32, N, M]	(3, 5)	1	(1, 2)

Each convolutional layer is followed by a Leaky ReLU activation with a slope of -0.2 and batch normalization, aiding in non-linear processing as well as data scaling. The weights are initialized using the Kaiming initialization, which is suitable for layers with ReLU activations.

Our NST model specifically aims to adapt the accentual style from an accent that SOTA SR models are typically trained on (style spectrogram) while preserving the original spoken content of input containing the non-standard accent (content spectrogram). Both spectrograms are passed through the CNN a single time to extract their respective features. Importantly, the weights of the CNN are fixed post-initialization, which ensures that feature extraction is consistent across iterations.

Following the feature extraction of the two input spectrograms, the model then initializes a generated spectrogram as random noise. This generated spectrogram is then refined iteratively through the CNN using the same fixed weights. Throughout this process, the model uses two distinct loss functions (content and style loss) to help update the generated spectrogram to ensure the structural elements align with the content input, and the textual features align with the style input. The optimization of the generated spectrogram is executed using the Adam optimizer, which iteratively adjusts the generated spectrogram to minimize the combined losses, blending the desired content structure with style features. Further details on the specific formulations of the content and style losses as well as the optimization strategy will be explored in the next section. The model's architecture and operational dynamics are visually summarized in Fig. 2.

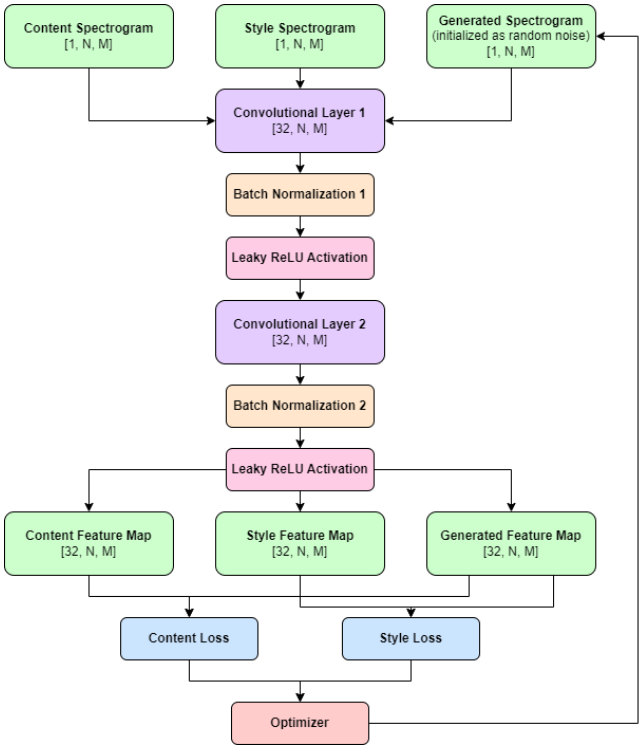

Fig. 2. NST Model Architecture

D. Loss Functions and Optimization

The objective of our NST model is to generate a new spectrogram that preserves the content of an input audio while altering its style to resemble that of an audio style reference. Thus, specific loss functions were used to measure the disparity between the content and style features of the two inputs. The model leverages two primary types of loss: content loss and style loss.

Content loss ensures that the generated spectrogram retains the original content spectrogram's structure and is defined as the Mean Squared Error (MSE) between the feature maps of the content spectrogram and the generated spectrogram. The loss is computed as follows:

$$L_{\text{content}} = \frac{1}{4 \times n_c \times n_H \times n_W} \sum (f_c - f_G)^2 \tag{1}$$

where m is the batch size, n_C, n_H, and n_W are the number of channels, height, and width of the tensor respectively, f_C and f_G are the feature maps of the content and generated spectrograms respectively.

Style loss, on the other hand, is concerned with capturing and imposing the textural features of the style spectrogram onto the generated spectrogram. It is quantified using the Gram matrix, which is a measure of feature correlation within a layer's feature maps. The style loss for a single layer is computed as:

$$L_{\text{style}} = \frac{1}{4 \times (n_c)^2 \times (n_H \times n_W)^2} \sum (G_S - G_G)^2 \tag{2}$$

where G_S and G_G are the Gram matrices of the style and generated feature maps, respectively. The Gram matrices are computed by reshaping the feature maps f from size (1, n_C, n_H, n_W) to (n_C, $n_H \times n_W$) and then calculating $GA = A \cdot A^T$.

The total loss used to train the NST model is a weighted combination of the content and style losses:

$$L_{\text{total}} = \alpha \times L_{\text{content}} + \beta \times L_{\text{style}} \tag{3}$$

where α and β are the content and style weighting factors, respectively.

E. Baseline Model Comparison

This study employs the Python Speech Recognition library as the foundational framework for the automatic speech recognition (ASR) model to establish a baseline for comparative analysis. This approach enables the integration of various ASR technologies and extends compatibility with different audio processing systems. The primary objective is to quantify the effectiveness of the proposed model in improving the understandability of ASR systems when applied to accent-modified audio files.

The evaluation of the style transfer modifications to audio files of diverse English accents is conducted with two key metrics. First, we examine the calculation of the Word Error Rate (WER) for both the original accented audio samples and the transformed audio samples against the ground truth transcriptions. The WER is given by:

$$WER = \frac{S + D + I}{N} \tag{4}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the number of words in the original transcription. By conducting this analysis we directly compare the WER of the original audio to the WER of the style-transformed audio, thus assessing the impact of our model. This comparison not only benchmarks the baseline performance of conventional ASR systems with audio inputs of various accents, but also quantifies the improvement in transcription accuracy achieved through our accent style transfer technique.

Second, we also examine the confidence level returned by the ASR system. This is an aggregate measure of the likelihood values assigned to each word used in the transcription. A higher level of confidence indicates that the system perceives less ambiguity in the speech sample, and thus we also seek to increase this value through the style transfer.

IV. Experimental Design and Hyperparameter Selection

Our study employs a fully automated pipeline designed to evaluate the performance of our NST model driven by a script that processes the audio files as specified in an input CSV file. This file contains the filename of the content input (the audio recording featuring a non-standard English accent), the specific segment of the recording to perform style transfer on, and the ground truth transcription for subsequent WER analysis. To ensure consistency across experiments, a single style audio input is employed universally for all instances of style transfer used in our experiment. It is identified by english494 in the Speech Accent Archive and represents a standard English accent on which prior work has shown commercial ASR systems to perform well [6].

The process for each audio sample in the input CSV involves several steps: both the content and style audio are trimmed to the designated segment, converted into their spectrogram representation, and then processed through the CNN for feature extraction. The generated spectrogram is run through the CNN for 5,000 epochs, optimizing the combined content and style loss using the Adam optimizer to refine the spectrogram output. Upon completion of the training phase, the generated spectrogram is transformed back into WAV format. The resultant audio is then analyzed to evaluate its ability to be understood by standard SR applications compared to the original content audio input. Using Python's SpeechRecognition library, we calculate the WER and confidence level for both the transformed and the original audio. These metrics are recorded in a results CSV file, providing a dataset for evaluating the NST's model performance across various accents. This approach allows for precise comparisons in style transfer effects and ensures the reproducibility of our experimental results.

Because our NST model trains on each content-style pairing of audio inputs, there is no notion of an explicit training split of the Speech Accent Archive dataset for our procedure. However, the selection of hyperparameters for the NST model was guided by preliminary experiments aimed at optimizing the training process. These experiments can therefore be considered the validation stage of our procedure and were conducted on a variety of accent types, including the Macedonian accent which is explored in greater qualitative detail later in this paper. The specific hyperparameters selected are shown in Table II.

TABLE II
Selected Hyperparameters

Number of Epochs	Content Weight	Style Weight	Optimizer	Learning Rate
5000	100	1	Adam	0.002

Initial tests with 20,000 and 10,000 epochs indicated that the model's loss metrics typically converge by around 5,000 epochs, which allowed us to reduce computational overhead by eliminating unnecessary training. The content and style weights influence the relative importance of the content versus style elimination in the loss function, which were set to 100 and 1 for content and style, respectively. This ratio ensures that the structure of the content audio is predominantly preserved, with stylistic elements introduced subtly. As outlined previously, the Adam optimizer was selected for its efficiency in handling sparse gradients and its adaptability in tuning the learning rates which is beneficial in handling unstable gradients encountered in our style transfer application. The learning rate is selected to be 0.002 to balance the speed of convergence as well as the stability of the optimization process.

For the purpose of testing our NST model, we defined a test split of the Speech Accent Archive dataset. This split consists of 10 randomly selected speech samples from speakers whose first language is each of Mandarin, Russian, and Spanish. In total, there are 30 speech samples in the test split. For each sample in this test set, we compare the WER of our baseline ASR model on the original sample against the transformed sample produced by our style transfer model. Due to time constraints imposed by the running time of the style transfer, we manually trimmed each pair in our test set such that both the content and style samples only included the opening phrases "Please call Stella. Ask her to bring these things with her from the store." The trimmed samples ranged from 4 to 9 seconds in duration.

The choice of speakers to use in our test set is motivated by prior work showing that out of all accent types presented in the Speech Accent Archive, commercial ASR systems tend to be least accurate when transcribing speakers whose first language is one of Mandarin, Russian, or Spanish [6]. We also ensured that no samples from these three languages were used during the validation stage of our procedure so that we would not manually tune the hyperparameters of our model according to its performance on the types of accents reserved for the testing stage.

V. Results

A. Quantitative Results

The results of our baseline ASR system on both the original and transformed audio for each sample in our test set are shown in Table III.

TABLE III
Model Performance Metrics

Accent	Original WER	Original Confidence	Transformed WER	Transformed Confidence
mandarin42	0.2857	0.7547	1.0	0.0
mandarin52	0.1429	0.8948	0.3571	0.8745
mandarin48	0.0714	0.7901	0.5	0.5290
mandarin23	0.2143	0.8907	0.2143	0.7389
mandarin34	0.1429	0.8489	1.0	0.0
mandarin45	0.0714	0.8738	1.0	0.0
mandarin16	0.3571	0.9251	0.5	0.7776
mandarin8	0.0714	0.7546	0.2857	0.5206
mandarin11	0.0714	0.9626	0.1429	0.8842
mandarin61	0.1429	0.9777	0.1429	0.7460
russian32	0.2143	0.8537	0.9286	0.8492
russian22	0.0714	0.9121	0.6429	0.8512
russian38	0.1429	0.8769	0.9286	0.1835
russian23	0.0	0.9221	1.0	0.0
russian19	0.0	0.9799	0.2143	0.9023
russian18	0.0	0.8436	0.1429	0.6319
russian1	0.0	0.9012	0.0	0.8289
russian9	0.1429	0.9220	0.1429	0.6983
russian24	0.1429	0.9193	1.0	0.0
russian17	0.1429	0.5207	1.0	0.0
spanish127	0.1429	0.9316	0.2857	0.9477
spanish142	0.0714	0.9702	0.0714	0.7765
spanish98	0.1429	0.9233	0.5714	0.8323
spanish140	0.2857	0.9059	0.4286	0.7319
spanish31	0.2857	0.9348	0.4286	0.6988
spanish67	0.1429	0.9567	0.2143	0.5842
spanish46	0.0714	0.9467	0.1429	0.7799
spanish1	0.2857	0.9317	0.2143	0.7127
spanish90	0.0714	0.9527	1.0	0.0
spanish143	0.0714	0.9661	0.2857	0.3550

On average, the transformed audios produced by our NST model yielded an average increase in WER of 0.3595 and an average decrease in confidence of 0.3436 against their corresponding original samples. Note that in cases where the WER on the transformed audio is exactly 1.0, it means that the baseline ASR was completely unable to understand the audio. Because these still represent failures in our method, we include such examples in our aforementioned computations of average changes in WER and confidence.

B. Qualitative Results

In addition to our experiments on the 30 short speech samples in the test set, we also report the result of our model on the speech sample of a speaker with a Macedonian accent (namely, macedonian19). This experiment was conducted during the validation stage of our procedure. As a result, it influenced our choice of hyperparameters and cannot be considered an unseen test example. However, it is unique in that it was the only experiment for which we had time to both learn the entire style sample and transform the entire content sample, rather than only the first two sentences of each. As such, we present this purely as a qualitative result rather than a true test of the model. Both the full macedonian19 content and english494 style samples used for this experiment were just over 20 seconds long.

On this pairing, the model was trained for a total of 10,000 epochs and achieved a loss of 1.7e-5, as shown in Fig. 6. Note that this differs from the 5,000 epochs used for the experiments on the test set, as it was conducted earlier in our procedure. The model achieved convergence quickly, as the loss remained fairly constant after 2,000 epochs.

Fig. 3 and 4 show the spectrogram representation of the two input audio recordings respectively. The signals from the content and style spectrogram are slightly misaligned due to the fact that the two speakers spoke at different speeds. In addition, the generated spectrogram can be found in Fig. 3c. Visually, the generated spectrogram shares many similarities with both the content and style spectrograms. Its signal follows the same timesteps as the content spectrogram (i.e. the location of the vertical lines is identical in both). Furthermore, the speech pattern from the style spectrogram is reflected in the generated spectrogram. Listening to the audio of the generated spectrogram shows a successful accent transfer (the audio file can be accessed on the GitHub repository).

Fig. 3. Spectrogram of audio with a Macedonian accent

Fig. 4. Spectrogram of audio with a standard English accent

Fig. 5. Generated spectrogram from combining Fig. 3 and 4

In terms of performance, the result of this Macedonian sample is described in Table IV. The differences in transcription accuracy arise from a single word that is transcribed differently in each, bolded below:

- **Original:**
  "please call Stella ask her to bring these things with her from the store 6 poyntz of fresh snow peas 56 **laps** of blue cheese and maybe a snack for her brother Bob we also need a small plastic snake and a big toy frog for the kids she can scoop these things into three red bags and we will go meet her Wednesday at the train station"
- **Transformed:**
  "please call Stella ask her to bring these things with her from the store 6 poyntz of fresh snow peas 56 **slabs** of blue cheese and maybe a snack for her brother Bob we also need a small plastic snake and a big toy frog for the kids she can scoop these things into three red bags and we will go meet her Wednesday at the train station"

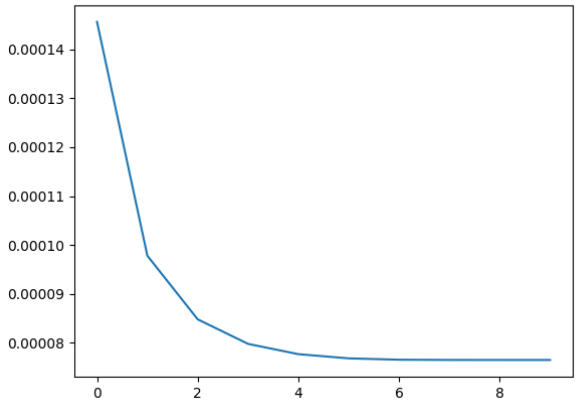

Fig. 6. Training loss of the generated spectrogram. The x-axis represents the number of epochs (in thousands) and the y-axis represents the sum of content and style loss.

TABLE IV
Model Performance on macedonian19

Original WER	Original Confidence	Transformed WER	Transformed Confidence
0.0725	0.9185	0.0580	0.8851

VI. Discussion

The empirical results of our NST model on the 30 samples in the test set are clearly disappointing. As detailed above, the transformations applied by the style transfer on the original samples yielded, in aggregate, significant increases in WER and decreases in confidence when passed to the baseline ASR system. Both of these trends represent a regression in our key metrics. We attribute this to the fact that the transformed speech sample produced by our NST model, when assessed on a qualitative basis, contains significant distortion due to the effect of blending the styles of two voices together. Indeed, there are many examples in the test set where the style transfer does not merely cause the baseline ASR system to misidentify several words, but instead be unable to recognize any speech in the audio entirely because it no longer sounds like human speech.

One common source of distortion that we identified arises from differences in the vocal range of the reference and content speaker. Specifically, because the reference sample used for all the experiments was spoken in a low, presumably male voice, applying its style to content samples spoken in higher, presumably female voices can result in highly noticeable audio artifacts. However, we maintained having a universal reference sample due to the risk of introducing biases by manually inferring the reference sample to pair with each content sample.

Despite this poor performance, it is important to note that all 30 samples in the test set were trimmed to be very short. In particular, both the content and style audio for every pairing that we experimented on were trimmed such that only the first two sentences of the ground truth text were recited. Although this was done for the purpose of reducing the computational cost of running our experiment, we recognize that we have potentially sacrificed accuracy in doing so. In particular, the trimmed samples contain only a fraction of all the English phonemes captured in the entire text, which limits our NST model from demonstrating its full potential in standardizing all possible elements of the content speaker's pronunciation style. Additionally, because the reference sample is also very short, it does not contain enough audio information for the model to adequately learn the reference speaker's style and distinguish it from random noise and idiosyncrasies in the sample that are unrelated to the speaker's accent. For all these reasons, we expect the model to yield more favorable results when the durations of both the content and style samples are increased.

Indeed, when testing the NST model on the entire macedonian19 sample, passing the transformed audio to the baseline ASR system yields a slightly lower WER than passing the original sample. This difference arises only from a disagreement in the transcriptions on one word, so this result is not intended to be considered rigorously. However, we interpret this as a promising sign that our NST model can viably augment an audio sample to improve its recognizability to ASR models when trained on long enough audio samples.

VII. Ethical Considerations

Although our proposal seeks to improve ASR accessibility, we acknowledge that implementing accent style transfer can inadvertently raise ethical concerns. First and foremost, our model defines one of the input audio samples to be a reference, the style of which is transferred to the content sample. The need to transfer accent styles may therefore imply that some accents or vernaculars are more valid and accepted than others, which is a harmful message that runs contrary to the principles of inclusivity and cultural diversity underlying this project. In practice, it may also introduce a new avenue of discrimination based on how the reference is used; for instance, a reference sample spoken by a male speaker is more likely to yield distortions when its style is applied to the content of a female speaker, due to general differences in their vocal ranges. Indeed, the experimental design in this paper falls victim to sending this message; we have consistently used audio samples from speakers with standard English or American accents as the style reference, and transferred such accents to the samples of speakers with generally non-standard accents. While this was done purely in the interest of improving model parity and accessibility, we acknowledge that our methods can cause users and developers of ASR models to feel uncomfortable and marginalized. We also acknowledge that, depending on how our methods are used, it may actually reinforce the issue of poor representation of uncommon accents in conventional speech datasets. This is especially the case if developers of ASR systems perceive accent style transfer to be a complete replacement for collecting more diverse and representative training data, rather than a supplementary tool.

Moreover, the potential misuse of our technology to alter voices raises ethical concerns surrounding privacy and identity theft. It is possible for malicious actors to use our methods to impersonate others by transferring the style of their speech to new content samples without the reference speaker's consent. Additionally, biases present in the training data could restrict the system's effectiveness across diverse accents. In particular, in our experimental design, we were only able to assess changes in the accuracy of ASR model performance across several accents already represented in the Speech Accent Archive, so this method does not completely guarantee the elimination of disparity between the countless speaking styles found worldwide. To address all of these concerns, we have committed to ethical data collection practices throughout our work on this project. We also emphasize that our model is intended only for alleviating disparities between the performance of ASR models, and that we have only demonstrated inconclusive improvements on a limited set of accents.

VIII. Conclusion

In conclusion, we have proposed a Neural Style Transfer model to perform the task of accent style transfer for the purpose of improving the inclusivity and accuracy of state-of-the-art ASR systems. Our proposed model takes a content speech sample spoken in a non-standard English accent and transforms its pronunciation characteristics by applying the style learned from a reference sample spoken in a standard English accent. At a qualitative level, we have established that this method indeed applies an appropriate style transfer via a comparison of the original and transformed spectrograms of samples spoken with non-standard accents. We have also shown that in a few instances, this method is able to improve the WER of a commercial ASR system when comparing the transformed audio against the original audio as a baseline. However, this was only possible on longer samples. On a test set of short samples, our experimental results generally showed an increase in WER and a decrease in ASR confidence largely due to the audio distortions produced by our model as well as the inherent complexity of distinguishing accent-related elements of a speaker's style.

Despite these results, we remain optimistic that with further computational resources and refinements to our model, the accent style transfer technique can be a viable method for improving parity in the performance of ASR systems on different accent types. We are most interested in improving the efficiency of the model so that it can more reasonably be applied to longer speech samples, as well as further experimenting with the model's hyperparameters and architecture to better learn the reference speaker's accent rather than other idiosyncratic features in the reference sample. Given the existing disparities in ASR performance shown in the literature, as well as the growing importance of speech recognition in human-computer interfaces, we emphasize that research in this direction is essential for creating more equitable and inclusive technologies that empower all users to communicate effectively.

References

[1] A. Aksënova, Z. Chen, C.-C. Chiu, D. van Esch, P. Golik, W. Han, L. King, B. Ramabhadran, A. Rosenberg, S. Schwartz, and G. Wang, "Accented Speech Recognition: Benchmarking, Pre-training, and Diverse Data," 2022, arXiv preprint arXiv:2205.08014. [Online]. Available: https://arxiv.org/abs/2205.08014

[2] A. Koenecke et al., "Racial disparities in automated speech recognition," Proceedings of the National Academy of Sciences, vol. 117, no. 14, pp. 7684–7689, Mar. 2020. doi:10.1073/pnas.1915768117

[3] A. Mani, S. Palaskar, and S. Konam, "Towards understanding ASR error correction for medical conversations," Proceedings of the First Workshop on Natural Language Processing for Medical Conversations, 2020. doi:10.18653/v1/2020.nlpmc-1.2

[4] L. A. Gatys, A. S. Ecker, and M. Bethge, "A Neural Algorithm of Artistic Style," 2015, arXiv preprint arXiv:1508.06576. [Online]. Available: https://arxiv.org/abs/1508.06576

[5] K. Fang, "randomCNN-voice-transfer: Implementation of Randomly Weighted CNN for Voice Style Transfer," 2023. [Online]. Available: https://github.com/mazzzystar/randomCNN-voice-transfer

[6] A. DiChristofano, H. Shuster, S. Chandra, and N. Patwari, "Performance Disparities Between Accents in Automatic Speech Recognition," 2023. [Online]. Available: https://openreview.net/forum?id=oqSKdRyYO1g